МОДЕЛЬ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ НЕСТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ КОРПОРАТИВНОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Нина Хайрова, Наталья Шаронова

Аннотация: *В работе предлагается математическая модель отношений между областью интеллектуальной деятельности менеджера и достоверными глубинными знаниями, представленными в документах, поступающих в корпоративную информационную систему (КИС). Доказывается возможность использования математического аппарата алгебры конечных предикатов в качестве базового средства описания модели. В работе показана возможность использования данной модели для извлечения знаний, представленных терминологическими понятиями и отношениями между ними, из разноформатных текстов документов КИС, с их одновременным динамическим разбиением на персонифицированные локальные области знаний менеджеров корпорации. Описывается информационное и алгоритмическое обеспечение программной реализации модели, приводится расчет эффективности использования разработанной подсистемы.*

Ключевые слова*: алгебра конечных предикатов, извлечение знаний, динамическая классификация, неструктурированная информация, корпоративная информационная система.*

ACM Classification Keywords *: H.3.3 .Information Search and Retrieval*

Введение

В последние годы достигнуты большие успехи в развитии КИС, основным направлением повышения эффективности работы которых, становится разработка интеллектуальных моделей трансформации информации, доступной снаружи и внутри организации, в инновационные знания. На передний план выходят требования по добыче «новых» для данной отрасли знаний, обеспечивающих фирму конкурентным потенциалом для принятия управленческих решений в изменяющемся бизнес-сообществе. Такого рода знания, в большинстве своем, представлены в текущих неструктурированных или слабоструктурированных документах организации. При этом, несмотря на достижения Text Mining, задача извлечения знаний из неструктурированной текстовой информации по-прежнему остается актуальной.

Современное понимание КИС подразумевает совокупность различных программно-аппаратных платформ, универсальных и специализированных приложений, интегрированных в единую информационно однородную среду, позволяющую принимать оптимальные для конкретного предприятия управленческие решения в соответствии с формализованными методами и правилами менеджмента. Одним из основных направлений повышения эффективности работы корпоративной информационной системы сегодня становится разработка интеллектуальных моделей трансформации информации, поступающей в организацию, в инновационные знания, представляющие интеллектуальные активы компании.

Рассматривая современные КИС с точки зрения управления знаниями можно заметить, что основной акцент сегодня делается не на сохранении разрозненной информации, а на извлечении закономерностей и принципов, позволяющих решать производственные и бизнес задачи, т.е. осуществлять накопление знаний [Гаврилова, 2000], [Mancini, 2001] . Источником знаний, как правило, служат документы, поступающие в систему на обработку: корпоративные стандарты, методики, бизнес-правила и технологии, а также технологическая и трудовая документация, накопившаяся в процессе функционирования организации. При этом перед КИС ставится задача извлечь и накопить именно инновационные знания,

которые снабжают фирму конкурентным потенциалом, а не коренные, устоявшиеся и даже «старые» знания, которыми обладают все участники данной отрасли. Чаще всего подобные «новые» знания, в большинстве своем, заложены в текущих, ежедневно поступающих в организацию документах.

Таким образом, основной задачей повышения эффективности КИС становится разработка системы трансформации информации, доступной снаружи и внутри организации в интеллектуальные архивы компании, представляющие инновационные знания. И если для решения данной задачи при работе со структурированной информацией используются хорошо разработанные технологии и методы data mining, то существующие сегодня методы излечения знаний из неструктурированных полнотекстовых документов не позволяют достаточно качественно разрабатывать единое информационное пространство, представляющее модель знаний предметной области работы КИС [Carroll, 2005].

Описание математической модели

В качестве базовых средств модели используем средства алгебры конечных предикатов (АКП). Вводим универсум элементов U, включающий все возможные текстовые документы, поступающие в корпоративную информационную систему на обработку, а также понятия и объекты анализа рассматриваемой предметной области, специализированные словари, тезаурусы, отображающие специфику данной предметной области. С данным универсумом U связаны некоторые знания, относящиеся к персонифицированному интеллектуальному ресурсу менеджера, т.к. они включены во все возможные текстовые документы, поступающие на обработку (справки, выписки, отчеты, распоряжения, решения и т.д.).

Из элементов универсума в соответствии с конкретной задачей обработки информации образуются подмножества M_{1i}, M_{2i}, …, M_{mi}, на декартовых произведениях которых $M_{1i} \times M_{2i} \times \ldots \times M_{mi}$ определяются предикаты P_j, характеризующие работу системы.

Предикатом P, заданным на универсуме, будем называть любую функцию, отображающую множество элементов универсума в ноль (предикат тождественно ложный) или единицу (предикат тождественно истинный). Так как множество элементов универсума информационной системы корпорации является конечным, то и предикат P соответственно конечен.

Базисным для алгебры предикатов является предикат узнавания предмета a по переменной x_i , равный единице, в том случае, если x_i равен a и нулю в противном случае, где i — это любой элемент универсума [Бондаренко, 2007]:

$$x_i^a = \begin{cases} 1, \text{ если } x_i = a \\ 0, \text{ если } x_i \neq a \end{cases}, \ (1 \le i \le n) \qquad (1)$$

Алгебра предикатов полна в том смысле, что любой ее предикат можно представить в виде суперпозиции базисных операций, примененных к базисным элементам. В работе [Шабанов-Кушнаренко, 1986] показано, что на языке АКП могут быть описаны любые конечные отношения, поэтому любой другой математический аппарат, предназначенный для описания произвольных конечных отношений, в логическом смысле, обязательно будет эквивалентен алгебре конечных предикатов.

Для построения модели описания персонифицированной локальной области знаний менеджера вводится конечное множество документов $D= \{d_i\}$, $1 \le i \le n$, поступающих менеджеру на обработку, и конечное множество терминологических понятий $T^k=\{t^k_j\}$, $1 \le j \le m$, первоначально определяющих область деятельности k-го менеджера, и представляющих его интеллектуальный ресурс, который был апробирован, разработан и использован данным или вышестоящим менеджером.

Под документом понимается некоторая целостная единица информации, имеющая уникальный идентификатор, средства отображения и модификации. Под терминологическим понятием понимается совокупность суждений о каком-либо объекте, отражающая его сущность, и выделяющая предметы некоторого класса по общим и совокупным специфичным для данного класса признакам.

На декартовом произведении D x T множеств D и T^k вводится бинарный предикат персонификации интеллектуального ресурса компании P (d, t), однозначно задающий отношение между обрабатываемым документом и терминологическим понятием интеллектуального ресурса менеджера:

$$P(d_i, t_j^k) = \delta \mid \; d_i \in D,\; t_j^k \in T^k,\; \delta = \{0,1\}. \tag{2}$$

Классификатор последовательно анализирует все возможные пары из множества D x T^k, однозначно приравнивая любой предикат $P(d_i,t^k_j)$ к нулю или к единице. Значение δ является истинным, если терминологическое понятие рассматриваемого документа определяет персонифицированную область деятельности менеджера, и является ложным в противном случае. Рассматриваемый предикат P удовлетворяет постулату существования [Хайрова, 2003]: предикат $P(d_i,t^k_j)$ реально существует в том и только в том случае, если при повторной итерации любой пары (d, t) из множества D x T^k классификатор будет реагировать тем же ответом, что и в первый раз.

Используя результаты исследования [Шабанов-Кушнаренко, 1993], для предиката персонификации интеллектуального ресурса компании можно ввести два определения. Два документа d_V и d_W относятся к области знаний k-го менеджера (d_V , d_W)∈D тогда и только тогда, когда для ∀ t^k: $P(d_v, t^k) = P(d_w, t^k)$. В этом случае можно говорить о том, что два документа тождественны по отношению к области знаний менеджера, и писать $d_V \sim d_W$. Два терминологических понятия t^k_V и t^k_W относятся к области интеллектуального ресурса k-го менеджера (t^k_V, t^k_W) ∈ T^k тогда и только тогда, когда для ∀d: $P(d, t_V^k) = P(d, t_W^k)$. В этом случае можно говорить о том, что два терминологических понятия тождественны по отношению к персонифицированному интеллектуальному ресурсу k-го менеджера, и писать $t^k_V \sim t^k_W$.

Рассмотрев все возможные пары на декартовом произведении d x t^k, получаем отображение Ω, множества рассматриваемых документов D на множество терминологических понятий T^k персонифицированного интеллектуального ресурса менеджера. Отображение Ω можно представить в виде двудольного графа, верхнее множество вершин которого представляет множество рассматриваемых документов D, нижнее множество вершин — множество терминологических понятий T^k интеллектуального ресурса k-го менеджера, а дуги показывают истинность предиката персонифицированного интеллектуального ресурса компании $P(d_i, t^k_j)$=1. На декартовом квадрате D×D универсума U вводим предикат соответствия документов персонифицированному интеллектуальному ресурсу k-го менеджера:

$$G_1(d_1, d_2) = \forall t^k \in T^k (P(d_1, t^k) \sim P(d_2, t^k)). \tag{3}$$

А на декартовом квадрате $T^k \times T^k$ универсума U вводим предикат соответствия терминологических понятий персонифицированному интеллектуальному ресурсу k-го менеджера:

$$G_2(t_1^k, t_2^k) = \forall d \in D (P(d, t_1^k) \sim P(d, t_2^k)). \tag{4}$$

Предикаты G_1 и G_2, определяемые выражениями (3) и (4) однозначно определяются предикатом P, рефлексивны, транзитивны и симметричны, из чего следует, что они являются предикатами эквивалентности. Предикат (3) можно использовать для объективного определения отношения двух любых документов d_1 и d_2, принадлежащих множеству D, к одной области знаний менеджера. Если $G_1(d_1,d_2)=1$, то при любом терминологическом понятии $t^k \in T^k$: $P(d_1, t^k)= P(d_2, t^k)$, т.е. информация данных документов, выражаемая терминологическими понятиями из множества T^k, позволяет отнести их к области знаний k-го менеджера. В противном случае, если $G_1 (d_1,d_2)=0$, то найдется такое терминологическое понятие $t^k \in T^k$, для которого $P(d_1, t^k) \neq P(d_2, t^k)$. В этом случае информация, передаваемая документами d_1 , $d_2 \in D$, выражаемая терминологическими понятиями t^k множества T^k, не совпадает и, следовательно, документы нельзя отнести к области знаний одного менеджера.

Предикат (4) можно использовать для определения соответствия двух любых терминологических понятий, принадлежащих множеству T^k, персонифицированному интеллектуальному ресурсу одного менеджера. Действительно, если $G_2 (t^k_1, t^k_2)=1$, то $P(d,t^k_1)= P(d, t^k_2)$ для любого докумена $d \in D$, то есть существуют терминологические понятия t^k_1 и t^k_2, которые либо одновременно относятся к области знаний k-го менеджера, либо одновременно не относятся. Иначе говоря, во множестве документов D нет такого документа $d \in D$, который бы одновременно и относился и не относился к области знаний k-го менеджера, выражаемой терминологическими понятиями $t^k \in T^k$. Если же $G_2 (t^k_1, t^k_2)=0$, то найдется такой документ $d \in D$, для которого $P(d,t^k_1) \neq P(d, t^k_2)$. То есть либо документ d включает знания, выражаемые терминологическим понятием t^k_1 и не включает знания, выражаемые терминологическим понятием t^k_2, либо, наоборот, документ d, включает знания, выражаемые терминологическим понятием t^k_2 и не включает знания, выражаемые терминологическим понятием t^k_1. В обоих случаях терминологические понятия t^k_1.и t^k_2 будут относиться к персонифицированному интеллектуальному ресурсу разных менеджеров.

Очевидно, что документы, входящие в полученные классы эквивалентности, не могут быть тождественными по смыслу безотносительно от терминологических понятий, отображающих область знаний того или иного менеджера, они являются эквивалентными относительно персонифицированной области знаний менеджера. Используя предикаты (3, 4), можно определить разбиение множества D на слои документов, а множества T — на слои терминологических понятий персонифицированных интеллектуальных ресурсов менеджеров корпорации. Класс ϑ_c всех документов $d \in D$, относящихся к области знаний одного менеджера, включающий документ $c \in D$, можно выразить как:

$$\vartheta_c(d) = \forall t \in T(P(d,t) \sim P(c,t)). \quad (5)$$

Класс Λ_b всех терминологических понятий $t^k \in T$, относящихся к персонифицированному интеллектуальному ресурсу k-го менеджера, включающий терминологическое понятие $b \in T$, можно представить следующим образом:

$$\Lambda_b(t) = \forall d \in D(P(d,t) \sim P(d,b)). \quad (6)$$

Формулы (5, 6) выражают разбиение документов, поступающих на обработку, и терминологических понятий, относящихся к области интеллектуальных активов менеджеров корпораций, на классы эквивалентности через предикат персонифицированного интеллектуального ресурса компании (2), однозначно определяемый классификатором.

Пример реализации модели

В рассматриваемой модели предметными переменными, формирующими терминологические понятия и определяющими отношение документа к предметной области деятельности менеджера, являются: l — ключевое слово или словосочетание документа; u — значение универсальной иерархической классификации (УДК) и r — предметная рубрика рубрикатора, принятого для систематизации областей знаний потока научно-технической информации. Поскольку данные переменные отражают суть документа, назначение и взаимосвязь его составляющих, можно говорить о том, что они объективно определяют истинные и достоверные глубинные знания, представлены в документах.

Также в модели используется, основное для наших рассмотрений, понятие области знаний менеджера q. Под областью интеллектуальных знаний конкретного менеджера организации будем понимать нечетко определенную часть корпоративных знаний, используемую для решения стандартных управленческих задач данного менеджера. Как известно, современная КИС является многопользовательской, то есть одновременно используемой конечным множеством топ-менеджеров данной организации при решении различных управленческих задач, при этом области деятельности менеджеров далеко не всегда совпадают с принятыми предметными областями. Под областью интеллектуальных знаний менеджера будем понимать совокупность некоторых предметных областей знаний, являющихся информационным обеспечением управленческих задач конкретного менеджера и выделяемых в некоторый класс по определенным общим и совокупным специфичным для данного менеджера управленческим ситуациям.

Например, при описании информационного пространства десяти документов, относящихся к финансово-экономической деятельности предприятия, на этапах предлингистического анализа статистико-позиционными методами [Khairova, 2009] определяются: множество ключевых слов $L=\{l^i\}$, $1 \le i \le 14$, где l^1 = депозитные операции, l^2= депонирование, l^3= банк, l^4= вексель,…, l^{14}= цензовая стоимость; множество значений $U=\{u^i\}$, $1 \le i \le 5$, где u^1=336.22 Налоги. Сборы, u^2= 336.24 Таможенные пошлины, u^3= 336.71 Банковское дело. Банки, u^4 =336.763 Ценные бумаги. Активы, u^5= 336.764 Биржевые операции; множество значений рубрикатора $R=\{r^i\}$, $1 \le i \le 4$, где r^1= банки и банковская деятельность; r^2=рынок ценных бумаг; r^3= экономика фирмы; r^4= финансовая экономика. Множество областей интеллектуальных знаний менеджера, обусловленное множеством рассматриваемых управленческих ситуаций, $Q=\{q^i\}$, $1 \le i \le 18$, где q^1= банковские операции, q^2 = транзитная торговля, q^3 = депонирование, …, q^{18} = налогообложение.

Можно построить парадигматическую таблицу, отображающую связь между областью локализации деятельности менеджера q^i и предметными переменными l, u и r. Используя данную таблицу отношений, выразим область локализации интеллектуальной деятельности менеджера q через значения предметных переменных *r*, *l* и *u*:

$$r^1u^1l^{12}= q^1 ; r^4u^2l^{12}= q^2; r^1u^1l^{13}= q^3 ; r^1u^1l^{14}= q^4 ; r^1 u^3 l^1= q^5; r^2u^4l^1= q^6; r^1u^3l^2= q^7; r^1u^3l^3= q^8; r^1u^3l^4= q^9; r^2u^4l^4= q^{10}; r^2u^4l^5= q^{11}; r^2u^4l^6= q^{12}; r^2u^5l^7= =q^{13}; r^2u^5l^8= q^{14}; r^1u^3l^8= q^{15}; r^4u^2l^9= q^{16}; r^4u^2l^{10}= q^{17}; r^4u^2l^{11}= q^{18}. \quad (7)$$

Выполняем операцию почленной дизъюнкции как можно большего количества родственных равенств. Родственными равенствами будем называть такие равенства, из которых после выполнения над ними операции почленной дизъюнкции можно получить равенства с левой частью в виде логического произведения, каждый сомножитель которого зависит от одной предметной переменной [Бондаренко, 2007]. Введение почленной дизъюнкции с использованием родственных равенств обусловлено необходимостью получения локальных областей интеллектуальных знаний менеджера, определяемых именем конкретного менеджера. Такие области могут включать больше чем одно исчисляемое ограниченное количество рубрик и предметных областей исследований.

$$r^1u^1(l^{12} \vee l^{13} \vee l^{14}) = q^1 \vee q^3 \vee q^4; \quad r^4u^2(l^{12} \vee l^9 \vee l^{10} \vee l^{11}) = q^2 \vee q^{16} \vee q^{17} \vee q^{18}; \quad r^2u^5(l^7 \vee l^8) = q^{13} \vee q^{14}$$
$$r^1 u^3(l^1 \vee l^2 \vee l^3 \vee l^4 \vee l^8) = q^5 \vee q^7 \vee q^8 \vee q^9 \vee q^{15}; \quad r^2u^4(l^1 \vee l^4 \vee l^5 \vee l^6) = q^6 \vee q^{10} \vee q^{11} \vee q^{12}. \tag{8}$$

Формируя функцию перехода от предметной области интеллектуальных знаний q к локальной области исследования менеджера m, в профессиональную деятельность которого входит данная область исследования q, и переопределяя зависимость локальной области исследования менеджера m от переменных r, l, u, получаем предикат $P(r, l, u, m)$, описывающий связь локальной области исследования менеджера и переменных, объективно определяющих глубинные знания документа:

$$P(r, l, u, m) = m^1r^1u^1 (l^{12} \vee l^{13} \vee l^{14}) \vee m^1r^2u^5 (l^7 \vee l^8) \vee m^2r^4u^2(l^{12} \vee l^9 \vee l^{10} \vee l^{11}) \vee$$
$$\vee m^3 r^1 u^3(l^1 l^2 \vee l^3 \vee l^4 \vee l^8) \vee m^4r^2u^4(l^1 \vee l^4 \vee l^5 \vee l^6). \tag{9}$$

Данный предикат наглядно изображается в виде логической сети (рис. 1), которая является графической интерпретацией результата бинарной декомпозиции многоместного предиката.

Каждому полюсу логической сети ставится в соответствие своя предметная переменная модели. С каждым полюсом связана область изменения атрибута этого полюса. Любой полюс логической сети в каждый момент времени несет определенное знание о значении своего атрибута. Каждой ветви логической сети ставится в соответствие свое бинарное отношение модели, которое называется отношением этой ветви. Каждая ветвь соединяет два полюса, отвечающие тем предметным переменным, которые связываются отношением, соответствующим данной ветви.

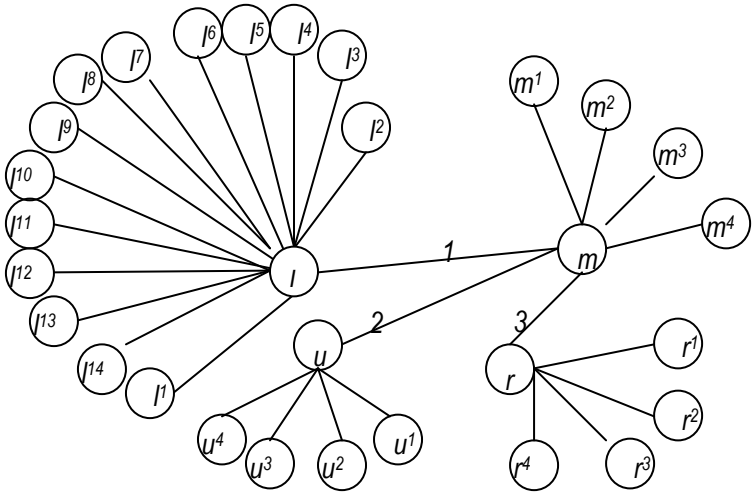

Рис. 1. Логическая сеть формального описания
локальной области деятельности менеджера

Программная реализация модели

Предложенный метод разбиения предметной области корпоративной информационной системы на локальные области деятельности менеджера реализуется в программном комплексе, включающем семь логических этапов, и представляющем семантически ориентированный лингвистический процессор: предлингвистическая обработка; графемная обработка; морфологическая обработка; контекстный анализ;

статистическая обработка; логико-алгебраическая обработка обучающей выборки; динамическая классификация поступающей информации.

На этапе графемной обработки выделяются элементы текста, имеющие графемное значение: заголовок, подзаголовок, абзац, предложение и т.д. Блок морфологической обработки вводится для учета словоизменительных форм и представления слов в канонической форме. Морфологический анализ осуществляется методом квазиокончаний. На этапе контекстного анализа из множества лексем выбираются словосочетания. Для учета информационной значимости ключевых слов на этапе статистической обработки вводятся весовые коэффициенты, являющиеся дополнительным средством семантической дифференциации лексических единиц документа. На этапе логико-алгебраической обработки обучающей выборки для описания связи предметной области деятельности менеджера, работающего с документами, с предметными переменными, объективно определяющих глубинные знания документа, используется полученное на предыдущих этапах работы системы информационное представление каждого документа в виде множества ключевых слов и словосочетаний, значений УДК и рубрик.

Строятся бинарные предикаты: $P_l(l,m)$, $P_r(r,m)$, и $P_u(u,m)$. Предикаты P_l, P_r и P_u можно представить в виде таблицы, заполняемой единицами или нулями в зависимости от значений соответствующих предикатов при данных значениях предметных переменных l, u, r, m. На практике часто встречаются ситуации, когда, исключая из рассмотрения некоторые элементы декартового произведения, можно получить разбиения множеств, более соответствующие интуитивным представлениям специалистов о семантике ключевых слов, предметных рубрик и значений УДК. Исключение некоторых элементов декартова произведения модели имеет смысл в том случае, когда упорядоченных пар не много по сравнению с общим числом элементов декартового произведения. Разработанный метод построения разбиений множеств *L*, *R* и *U* учитывает такие исключения. Алгоритм допускает небольшие различия между строками таблиц, попадающими в один класс. Мера таких допустимых отклонений ρ' может устанавливаться пользователем. При этом те строки (или столбцы), которые могут быть отнесены к различным классам, выделяются особо и могут быть классифицированы пользователем отдельно. Такие строки (столбцы) отличаются от элементов некоторых классов на ρ двоичных разрядов, не превышающем заданного ρ' ($\rho \le \rho'$), где ρ можно интерпретировать как расстояние между векторами: $\rho(a,b)=\Sigma\ \alpha_i \oplus \beta_i$; a и b — двоичные векторы: $a=(\alpha_1, \alpha_2, \ldots, \alpha_k)$, $b=(\beta_1, \beta_2, \ldots, \beta_k)$.

На последнем шаге этапа обучения полученные классы ключевых слов, УДК и значений рубрикатора вводятся в эталоны, представляющие множества $U_1, U_2, \ldots, U_n$, относящиеся к областям исследования конкретных менеджеров. Разбивая таким образом терминологические понятия обрабатываемых КИС документов, мы иерархически упорядочиваем их, используя этого отношения «быть элементом класса» и «включаться в предметную область исследования».

Для последующей динамической классификации множества документов, поступающих на вход системы, выполняются процедуры предлингвистической, графемной, морфологической, контекстной и статистической обработки, в результате которых каждому документу приписывается его информационное представление. Этап сравнения полученного информационного представления с имеющимися эталонами позволяет отнести документ к соответствующим областям деятельности того или иного менеджера. Для вновь полученных документов можно повторить этап обучения для создания обновленных эталонов множеств терминологических понятий $U_1, U_2, \ldots, U_n$, относящихся к областям исследования конкретных менеджеров. При этом система может гибко и динамично менять разбиение понятий и объектов, обрабатываемых менеджером в процессе решения управленческих задач, и, соответственно, менять систему классификации документов, поступающих на вход КИС.

Так как в настоящее время нет четких и однозначных определений понятий «эффективного излечения знаний» и «качества знаний», количественная оценка результатов работы системы не является тривиальной задачей. Для оценки эффективности работы системы используем метод тестовых коллекций [Cormack, 1998], заключающийся в сравнении выводов, сделанных системой, с мнением экспертов.

Экспертами, которыми являлись менеджеры корпорации, работающие с документами в КИС, определяется релевантность отнесения документов к области деятельности менеджера. Отношение релевантности является субъективным сложно определяемым, при определении эффективности работы системы, мы основывались на определении релевантности, используемом в [Mizzaro, 1996], согласно которому релевантность зависит от четырех понятий: Relevance (IR, IN, C, T), где IR — информационный ресурс, IN — информационная потребности, C — контекст и T — время. В предлагаемой модели информационный ресурс представлен множеством электронных документов, поступившим на обработку IR = D. Для получения интегральных показателей качества работы системы перевода информации в интеллектуальный активы компании применялись методики усредненных метрик [Кураленок, 2002]. В качестве метрик использовались коэффициенты точности, полноты и шума, а для вычисления среднего гармоничного точности и полноты использовалась мера Ван Ризбергена.

Исследовалась выборка из трехсот документов, поступающих в организацию, в среднем в течение десяти дней. Средний коэффициент полноты, определяемый отношением числа релевантных документов, отнесенных к области деятельности менеджера, к общему числу поступивших в систему документов, релевантных области деятельности данного менеджера, recall = 0,91. Средний коэффициент точности, определяемый отношением числа релевантных документов, отнесенных к области деятельности данного менеджера, к общему числу документов, отнесенных к области деятельности данного менеджера, precision = 0,825. Коэффициент шума, определяемый отношением числа нерелевантных документов, отнесенных к области деятельности менеджера, к общему числу документов, отнесенных к области деятельности данного менеджера, error = 0,03. Сбалансированная мера Ван Ризбергена была определена как F_1= 0,8654.

Сравнивая полученные показатели с коэффициентами полноты и точности аналогичных систем, следует отметить, что если значения коэффициента полноты современных гипертекстовых информационно-поисковых систем в среднем recall $\geq$0,9, а точности precision $\geq$0,8, то значения данных коэффициентов для документальных систем значительно ниже, в среднем $\approx$ 0,7 [Башмаков, 2005]. Кроме того, современные системы классификации узкоспециализированных полнотекстовых документов, используемые в отраслях, применяют статические системы классификации, со строго заданными иерархическими отношениями между тематическими классами. В случае расширения предметных областей и развития знаний, точность и полнота выдачи таких систем резко падает и требует ручного изменения классификационной схемы [Бабенко, 2004].

Выводы

Результатом данного исследования является математическая модель отношений между областью интеллектуальной деятельности менеджера и достоверными глубинными знаниями, представленными в документах. Реализация модели дает возможность осуществлять динамическое разбиение поступающих в КИС документов и содержащихся в них терминологических понятий. Использование данной модели позволило разработать информационное и программно-алгоритмическое обеспечение семантико-ориентированного лингвистического процессора, основанного на работе логических сетей, извлекающего новые терминологические понятия из поступающих в систему документов и относящего их к классам, соответствующим областям деятельности различных менеджеров. Извлекаемые понятия наполняют новой информацией базы знаний корпоративной информационной системы, делая ее динамичной и одновременно увеличивая семантическую силу модели представления знаний КИС. Приведенные расчеты эффективности программной реализации модели показывают достаточно высокую полноту и релевантность системы разбиения документационных потоков по соответствующим областям деятельности менеджеров КИС.

Литература:

[Carroll, 2005] Carroll, J., Evans, R., and Klein, E. Supporting Text Mining for e-Science: the challenges for Grid-enabled Natural Language Processing. UK e-Science All Hands Meeting, Notingham, UK, 2005.

[Cormack, 1998] Cormack G.V., Palmer C.R., Clarke C.L.A., Efficient construction of large test collections // Proc. of the SIGIR'98 - pp. 282-289.

[Khairova, 2009] Nina Khairova, Natalia Sharonova. Use of Predicate Categories for Modelling of Operation of the Semantic Analyzer of the Linguistic Processor./Proceedinga of IEEE EAST-West Design & Test Symposium (EWDTS'09). // Moscow, Russia, September 18-21, 2009/ — P. 204- 207

[Mancini, 2001] Mancini J. Enterprise Content Management: Critical Technologies for Business Applications // AIIM, 2001.

[Mizzaro, 1996] Mizzaro S. How many relevances in information retrieval?, in C.W. Johnson and M. Dunlop (eds), Proceeding of the Workshop 'Information Retrieval and Human Computer Interaction', GIST Technical Report GR96-2, Glasgow University, The British Computer Society, Glasgow, UK, pp. 57-60.

[Бабенко, 2004] Система классификации текстов информационных сообщений на русском языке "АКТИС"/ Бабенко М. и др. // Труды международной конференции "Программные системы: теория и приложения", ИПС РАН, г. Переславль-Залесский, май 2004 — М.: Физматлит, 2004, Том 2., с. 7-21.

[Башмаков, 2005] Башмаков А. И., Башмаков И. А. Интеллектуальные информационные технологии. — М. : Изд-во МГТУ им. Н.Э.Баумана, 2005. — 304 с.

[Бондаренко, 2007] Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П. Теория интеллекта. Харьков. Изд-во СМИТ. 2007. 576 с

[Гаврилова, 2000] Гаврилова Т. А. Базы знаний интеллектуальных систем. – СПб : Питер, 2000.

[Кураленок, 2002] Кураленок И., Некрестьянов И., Оценка систем текстового поиска. // Программирование. 2002, 28(4):226-242.

[Хайрова, 2003] Хайрова Н.Ф., Шаронова Н.В. Автоматизированные информационные библиотечные системы: задачи обработки информации: Монография. — Х.: Нар укр. акад., 2003. – 120 с.

[Шабанов-Кушнаренко, 1986] Шабанов-Кушнаренко Ю.П. Теория интеллекта: Технические средства. — Х.: Вища школа, 1986.— 134с.

[Шабанов-Кушнаренко, 1993] Шабанов-Кушнаренко Ю.П., Шаронова Н.В. Компараторная идентификация лингвистических объектов: Монография.— К.: ИСДО, 1993.— 116 с.

Сведения об авторах

Хайрова Нина *– доцент кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический университет», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail: nina_khajrova@yahoo.com*

Научные интересы: искусственный интеллект, обработка знаний, автоматическая обработка текстов

Шаронова Наталья *– профессор, заведующий кафедрой интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический университет», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail:, nvsharonova@mail.ru*

Научные интересы: искусственный интеллект, математическое моделирование, автоматизированные библиотечные системы